Project Report

Computational Data Science
CMPT 353 D100
Instructor: Greg Baker

Jacky Lim - 301383034
Sam Wong - 301372150
Sanshray Thapa - 301456437

August 4, 2023

The Problem

Have you ever wondered if the suicide rate of a country is associated with their citizens' happiness? Do certain quality of life categories of a country such as generosity, freedom, and social support have an impact on citizens' likeliness to commit suicide? What is the greatest factor that causes someone to be happy or unstable? That is what we are trying to find out.

Initially, we were just interested in analysing different metrics of happiness like GDP, wealth, and freedom and seeing which one impacts suicide rates the most. As we continued with the project, we also decided to analyze the suicides data by itself. We wanted to know how they differed for various metrics such as age, region, and gender. Out of these metrics, we wanted to go more in depth on regions. We wanted to know how each region differs from each other in terms of the metrics of happiness listed above.

The Data

We used two datasets to conduct our analysis on the correlation between suicide rates and different metrics of happiness: world happiness data from the World Happiness Report (https://worldhappiness.report/ed/2023/#appendices-and-data) and mortality data from the World Health Organization (https://platform.who.int/mortality/themes/theme-details/topics/indicator-groups/indicator-group-details/MDB/self-inflicted-injuries).

First Look at the Data

We first began by looking at the world happiness data to see the different metrics of happiness, and how much we needed to clean data. We looked at how many columns have missing data and printed some of them. We looked at the number of countries, as well as the number of entries for each year. We saw that the earlier years, like the 1950s, had very few rows. We knew that the mortality dataset didn't have data from that time period, so we planned to drop them when merging. Lastly, we checked the normality of each metric and saw that all columns had a p-value that was a lot smaller than 0.05 after a normality test. We then concluded that linear regression would be the best statistical tests to use for future analysis as other statistical tests like the t-test wouldn't be accurate even if we were to transform the functions.

Combining and Cleaning the Data

To combine the data to analyse how different metrics of happiness affect suicide rates, we merged the two datasets together by matching rows containing the same country name and year. This process removed any row that did not have a matching counterpart from one dataset to another. After combining the two datasets, we removed columns we deemed unnecessary like duplicates of country name and year we got from merging. The mortality dataset split the data into different age groups and sex. We only kept the rows containing data for all age groups and sex. We then removed all rows with missing data. Examining the combined dataset, we

recognized that all 0 values should be treated as missing values. As such, we removed all rows containing 0 values as well. Finally, we saved this new combined and cleaned dataset in a csv file to be used for analysis.

Metrics of Happiness

We mentioned the different metrics of happiness that we will be using to analyse their correlation to suicide rates. Some of these metrics were self-explanatory, however, others required further explanation in order to understand them. The following is a list of the different metrics of happiness, their definitions, and how they were measured. Some of the metrics get their value based on the national average of a specific question asked in the Gallup World Poll. The question has a numerical answer (0/1 for yes/no), so they range between 0 and 1.

Happiness score / ladder rank:
- Rating out of 10 on citizens' happiness

Log GDP per capita:
- The logarithm of a GDP divided by the population

Social support: Average of the responses (yes/no) to the question:
- "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"

Healthy life expectancy at birth:
- The average number of years a newborn can expect to live in good health

Freedom: Average of the responses (yes/no) to the question:
- "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"

Generosity: Average of the responses (yes/no) to the question:
- "Have you donated money to a charity in the past month?"

Perception on corruption: Average of the responses (yes/no) to the questions:
- "Is corruption widespread throughout the government or not?"
- "Is corruption widespread within businesses or not?"

Positive affect: Average of the responses (yes/no) to the questions:
- Did you experience laughter and/or enjoyment during a lot of the day yesterday?
- Did you do something interesting yesterday?

Negative affect: Average of the responses (yes/no) to the questions:
- Did you experience worry, sadness and/or during a lot of the day yesterday?

Additional details can be found on Technical Box 2:
https://worldhappiness.report/ed/2023/world-happiness-trust-and-social-connections-in-times-of-crisis/#ranking-of-happiness-2020-2022

Techniques Used in Analysis

Simple Linear Regression

We used simple linear regression to determine if there is any correlation between the metrics of happiness and suicide rates. Simple linear regression or ordinary least squares test is a statistical test used to see if a variable y is linearly dependent to variable x. In our case, y is suicide rates and x are the different metrics of happiness. Our null hypothesis was that suicide rates does not linearly depend on the different metrics of happiness; our alternative hypothesis was that it does. At a 95% confidence interval, we would need a p-value of less than 0.05 to reject the null hypothesis.

Checking Normality of Residuals

Simple linear regression or ordinary least squares assumes that the residuals are normally distributed. The residuals is the difference between the observed data and the fitted data. The central limit theorem states that when sampling n values, the means of samples are normally distributed for a large enough n. In practice, that n is usually greater than or equal to 40. Based on the data in our analysis for linear regression, we have enough data points to assume normality as long as the distribution of residuals looks about normal. Looking at the histograms of residuals in Figure 1 below, we are able to see that the residuals are normal enough meaning we can assume normality.

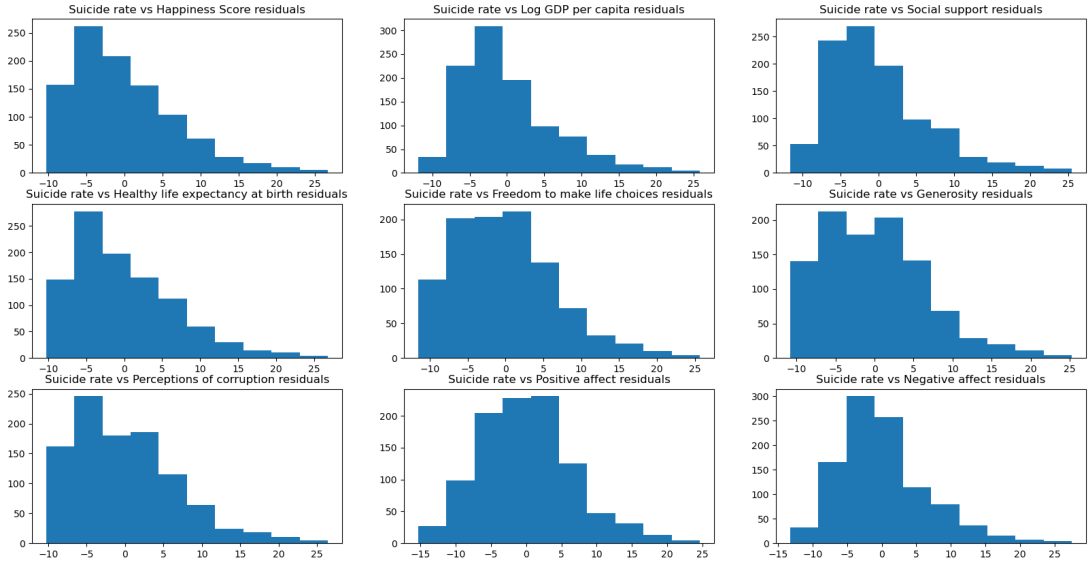

Figure 1 - Histograms of residuals for suicide rate vs metrics of happiness

Random Forest Model

Because of the large dataset with multiple dimensions, we continued our analysis by creating a RandomForest model. We used RandomForestRegressor to predict Happiness Scores and Suicide Death Rates per 100,000 population, while we used RandomForestClassifier to predict the region name. The main benefit of creating this model is to create the highest accuracy prediction and analysed the highest ranked metrics features like 'Log GDP per capita', 'Social support', 'Freedom to make life choices', and others, in predicting happiness scores, suicide rates, and region classification.

Results

Normality and Equal Variance Tests

Before performing the regression analyses and calculating the below, normal and equal variance tests were performed. What we found was that none of the data was equally distributed or had equal variance. As a result, we used a non-parametric version of ANOVA (Kruskal-Wallis Test) and found p-value less than 0.05. This indicates there are significant differences in the distribution. Example:

Metric	stats.normaltest	stats.levene	p-value
Happiness Score	4.86e-05	6.81e-21	2.10e-30
Log GDP per capita	1.30e-07	3.24e-23	3.03e-64
Positive affect	6.79e-15	1.68e-45	1.70e-78
Death rate	1.57e-25	7.37e-15	2.74e-65

Linear Regression

As stated in the Techniques Used in Analysis section above, we used linear regression to attempt to answer our initial question on whether or not the different metrics of happiness impacts suicide rates. Our null hypothesis was that suicide rates does not linearly depend on the different metrics of happiness, or alternatively, that the slope line is 0. Conversely, our alternative hypothesis was that suicide rates does linearly depend on the different metrics of happiness.

The following are the p-values and correlation coefficients we got from running linear regression on suicide rates against the metrics of happiness:

	p-value	correlation coef
Happiness Score	1.387833e-02	0.077436
Log GDP per capita	3.935033e-24	0.311428
Social support	2.859620e-20	0.284710
Healthy life expectancy at birth	3.809343e-03	0.091016
Freedom to make life choices	3.635846e-02	-0.065897
Generosity	4.276620e-05	-0.128447
Perceptions of corruption	8.766034e-01	0.004894
Positive affect	3.402680e-11	-0.206659
Negative affect	2.796164e-25	-0.318866

As seen in the results above, all metrics of happiness other than perceptions of corruption have a p-value less than 0.05. This means at a 95% confidence interval, we are able to reject the null hypothesis for the metrics of happiness aside from perceptions of corruption. Based on this result, we can say that suicide rates is linearly dependent on the different metrics of happiness other than perceptions of corruption. For perceptions of corruption, we can accept the null hypothesis at a 95% confidence interval as its p-value is roughly 0.88 which is greater than 0.05. This means that suicide rates does not linearly depend on the perceptions of corruption.

Looking at the correlation coefficient (r), we are able to see that although the p-values show significance between the metrics of happiness and suicide rates, the correlation is rather low for almost all metrics. Correlation coefficients span from -1 to 1. Values closer to -1 and 1 signify a higher correlation and values closer to 0 signify a lower correlation, with negative values signifying a negative correlation. In our results, the metric with the highest correlation coefficient was negative affect at around -0.319 followed closely by log GDP per capita at around 0.311. Although they had the highest correlation coefficient out of the metrics of happiness, we cannot conclude that they have any correlation with suicide rates as 0.3 is still rather low. The metric with the lowest correlation coefficient was perceptions of corruption at 0.005. We can safely conclude that perceptions of corruption does not have any correlation with suicide rates.

Based on the results of linear regression on suicide rates against the metrics of happiness, we are unable to conclude that the different metrics of happiness have any significant linear correlation with suicide rates. Although the p-values suggest that there is a relationship between suicide rates and the metrics of happiness, analysing the correlation coefficients showed that the correlation between the variables are rather low. However, we can safely conclude that the perceptions of corruption does not have any linear correlation with suicide rates as its p-value was higher than 0.05 allowing us to accept the null hypothesis, and its correlation coefficient was extremely close to 0 signifying low to no correlation.

To help visualise our findings, Figure 2 shows the graphs of suicide rates versus the metrics of happiness with the original data points in blue and the best fit line in red. We can clearly see that the slope of the best fit line for the suicide rates versus perceptions of corruption graph is almost 0. This aligns with us accepting the null hypothesis as we stated that the null hypothesis can also be interpreted as the slope line being 0. We can also see that the rest of the graphs show no clear indication of a linear trend meaning that our inconclusive results on the rest of the metrics of happiness are understandable.

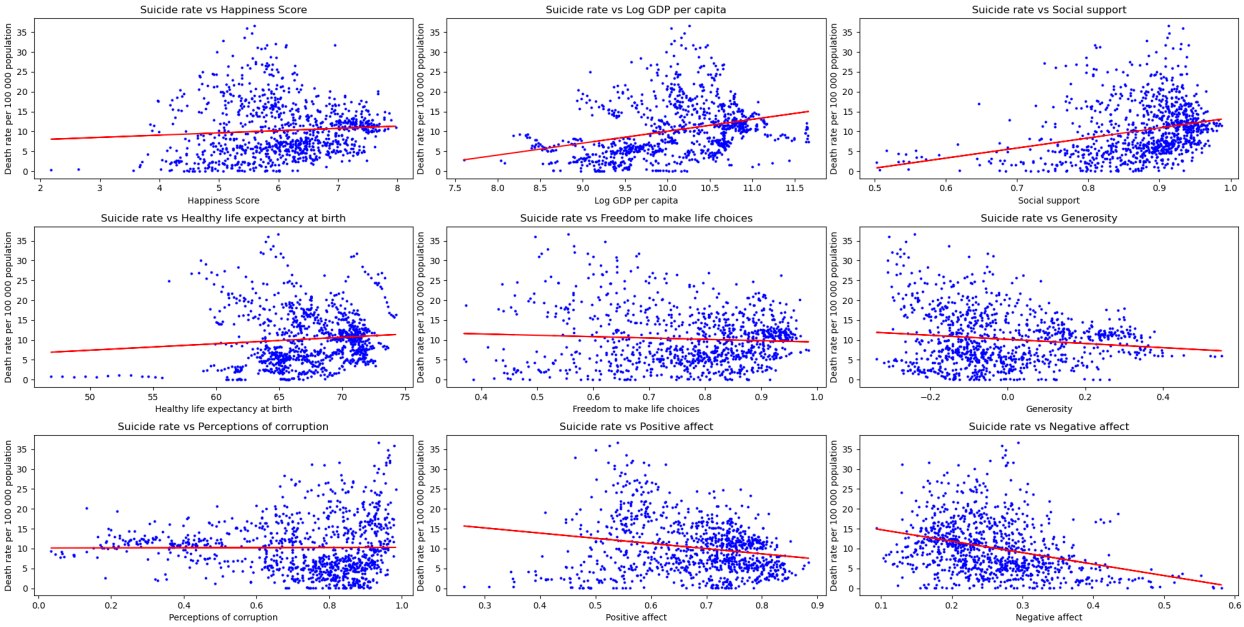

Figure 2: Linear regression plots for suicide rate vs metrics of happiness

Correlation Heatmap

Because of the large amount of metrics, creating a heatmap helps us easily visualise the correlation between multiple different variables. The dark red/blue represent strong correlation, while the whiter spots indicate low correlation.

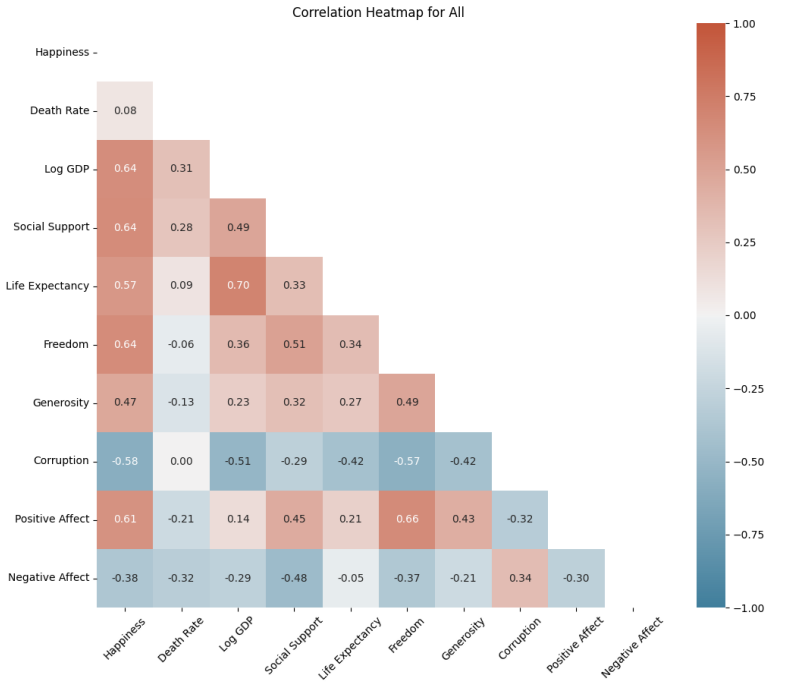

Figure 3: Correlation Heatmap

From the heatmap in Figure 3, we can see that most metrics have a strong correlation with one another. Most notably, GDP has a large correlation with Life Expectancy and Social Support, and perceptions of corruption have a significant negative correlation with happiness, GDP, and perceptions of freedom. This seems to confirm wealthier countries tend to have better support systems and less corruption, resulting in a happier society. It is worth noting that the suicide rate has a low correlation with GDP, Social Support, Positive/Negative Affect.

Models

For all the models, RandomForest with 125 trees was used. We found that this yielded the highest score. For Happiness, the average scores after 10 runs is 0.84, the scores for predicting suicide rate was 0.65, and the scores for predicting the region was 0.87.

Happiness: The most important feature for predicting Happiness by far was GDP. Positive Affect was the 2nd most important thing. Having positive emotions such as laughter, enjoyment, and doing or learning something interesting greatly has a profound impact on happiness scores. This shows the relationship between economic prosperity and happiness and also between financial struggles and increased mental suffering. Everything else has insignificant impact.

Death Rate: GDP was also the most important feature for predicting the suicide rate followed by Positive Affect. Everything else has insignificant impact.

Region Classifier: When trying to predict the region, GDP and Positive was the most important but significantly less than predicting Happiness and the Death Rate. Other metrics such as Social Support, Generosity, and Freedom, were weighted significantly more than other models. We assume this stems from the various levels of freedom, social programs, and culture differ across the regions. These insights demonstrate the importance of having social support and how greatly they influence the well-being of its citizens.

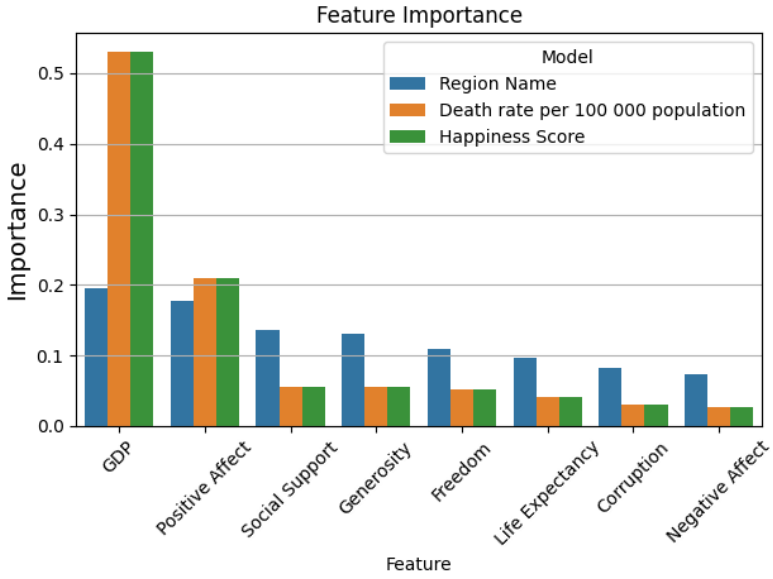

Figure 4: Model Feature Importance

Limitations

The first obstacle we ran into was to choose the right datasets for our project. After brainstorming, we decided that a good project idea is to find out if happiness and suicide are correlated, and to see how strongly they are related if so. If that were the case, we wanted to know what the different factors of happiness are and how do each of the contribute in the overall relationship of happiness and suicide rates. We looked at a couple of datasets for happiness report and mortality rate, and chose the datasets that were the most recent, and had the most metrics to choose from when analysing.

Additionally, we found some of the happiness metrics a little questionable, mainly how they were measured. For example, metrics such as freedom and generosity were measured by a singular yes/no question in a survey. A singular question that asks if a person donated does not feel like it is enough to determine if they are generous. The data could have been more accurate if there were more questions asked in the survey, perhaps asking questions to give a number from 1 to 10 on how much they agree or identify with a particular question would be better than a yes/no question. These types of questions would allow participants to express their opinions better, while still having data that can be represented numerically, and used for analysis.

Since the datasets we chose were big and contained 165 countries, they were bound to have missing data. We decided to make the major decision of dropping NaNs instead of imputing because we didn't want to introduce any bias. Also, we believe that with the large data set (over 1000 rows), dropping NaNs would still realistically represent our sample for analysis. Another issue with our dataset was the inconsistent availability of data. Some countries had missing data for certain years, or some of the metrics for that year were missing. This made it a challenge to map out trends accurately as it may introduce bias towards countries that report more data frequently. This influenced our decision to not analyse individual countries, and to find trends by grouping by regions to minimise any errors caused by dropping NaNs, and the small amount of data some countries had.

Another issue we ran into was trying not to overfit when we implemented RandomForestRegressor, and RandomForestClassification models to predict happiness scores, suicide rates, and regions. The challenging part was to find the right parameters to make the training and test data scores to be as close as possible to ensure we weren't overfitting. The average training dataset score for all three models was above 0.9. The average scores for the happiness and region model were not too far off the training dataset, but the suicide rate model score was nowhere close, even after trying various values for the parameters. We then came to a conclusion that we cannot use different metrics of happiness to predict suicide rates.

```
X = data[features]
y = data['Death rate per 100 000 population']

# Splitting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y)

# Train a Random Forest Regressor for "Happiness Score"
model = RandomForestRegressor(n_estimators=125, max_depth=50, min_samples_l
model.fit(X_train, y_train)

print(model.score(X_test, y_test))
print(model.score(X_train, y_train))
```

```
0.636684255250803
0.9528169073254453
```

Lastly, some things we could have done if we had more time are to try other regression and classification models. We could see how the average score of each test dataset differed when using other models such as kNN and neural network. When dealing with analysis that compared regions, we could have tried to plot it on a world map instead of the different graphs we used.

Project Experience Summary

Jacky

- Obtained and prepared datasets for cleaning and analysis, ensuring all data are labelled consistently. Ensure data across different datasets can be correctly merged.
- Conducted Normality Tests, Equal Variance test and Kruskal-Wallis test to determine if there were significant differences in the distribution.
- Developed a RandomForestRegressor model to predict variables such as suicide rates and happiness. Developed a RandomForestClassifier to predict the regions based on the metrics included in the data.
- Generated a heatmap to easily understand and visually compare the correlation between all the different metrics for deeper analysis using seaborn.

Sam

- Merged a dataset with 2000 rows of data with a dataset containing over 200,000 rows and cleaned the merged dataset using the Pandas library in Python to create a new dataset of 1000 rows to be used for analysis
- Performed linear regression on the merged and cleaned dataset using the scipy.stats Python library to analyse the relationship between suicide rates and different metrics of happiness
- Utilised the matplotlib Python library to produce plots to help visualise the results found through linear regression
- Analysed statistics found through linear regression by looking at the p-values and correlation coefficients to make conclusions on the data

Sanshray

- Analysed a world happiness dataset to determine what countries had missing or little data, and see what statistical tests would be best to perform
- Utilized the matplotlib and Pandas Python libraries to create insightful plots that show how suicide rates differ between various regions and age groups over time
- Observed results that were produced by the RandomForestRegressor, and RandomForestClassification models to determine if they can predict happiness, suicide rates, and region
- Explained the obstacles, limitations, and flaws with our analysis